Nicu Sebe
University of Trento, Italy
sebe@disi.unitn.it
An intelligent information system should always be user-centric!!!
Digital Natives?

Play video
What’s in my backpack
What is happening?

- Multi-cultural, multi-lingual environments, large (and instant) access-to and storage-of multimedia information (documents, sensors: RFID, etc.)

- A variety of devices (cell phones, meeting rooms, desktop systems) and media (voice, video, text) for access, different bandwidths

- Differences across time and space, lower communication costs, more asynchronous collaboration, annotated collections (communities and social networks).

*Interactive & Distributed Environments!*
Interactive Environments

computing these days is a lot about media and people

production, edition, annotation, organization,
archival, retrieval, sharing, analysis,
communication, …

for

learning, teaching, creating, caretaking,
communicating, working, playing, entertaining,
dating, childbearing, shopping,…

(and also for bombing, killing, …)
Human-Centered?

We experience this form of interaction often (as I was typing this font changed!)

Sometimes we experience problems with new media!!!!!

Boom Chicago of Amsterdam made the following video in anticipation of technological advances in US elections.

It can be seen in our theater as part of our show ‘Mr. America Contest’ or in voting booths across Florida Nov. 2, 2004...
Main problems

- We don’t think of real users (innovation driven by technology)
- Thinking of users is not enough
- Most technology development improves on existing technology (where human issues played marginal role)
- Huge impact on social, cultural, economic, political aspects of life!!!
What is HCC?

Shneiderman’s “new computing”
– Leonardo as muse
– how would Leonardo use a laptop and what applications would he create?
– shifts focus from what computers can do to what users can do
– support creativity, consensus-seeking, and conflict resolution

IEEE 2003 book award
Where is HCC?

Technology
- Algorithms
- Applications
- No “users”

Human centered
- Combination

User
- User requirements
- User studies
- Interaction
- ....
Human-Centered Computing

- Human Computer Interaction
- Content analysis
- Human issues
- Machine Learning, Personalization, Context, ...
- Computer vision
Characteristics of HCC

- Core of HCC system and algorithm design processes:
  - Socially and culturally-aware
  - Directly augment and/or consider human abilities
  - Adaptable

- HCC applications should exhibit the following qualities:
  - Act according to the social and cultural context in which they were deployed
  - Integrate input from different types of sensors and communicate through a combination of media as output
  - Allow access by a diversity of individuals.
HCC Computing Cluster Topics (NSF)

- Systems for problem-solving by people interacting in distributed environments.
  - Internet-based information systems
  - Sensor-based information networks
  - Mobile and wearable information appliances

- Multimodal interfaces used by people to communicate with one another

- Intelligent interfaces and user modeling, information visualization, and adaptation of content to accommodate different display capabilities, modalities, bandwidth, and latency
HCC Computing Cluster Topics (NSF)

- Multi-agent systems that control and coordinate actions and solve complex problems in distributed environments
  - e-Commerce
  - Medicine
  - Education

- Models for effective computer-mediated human-human interaction under a variety of constraints
  - Video conferencing
  - Collaboration across networks
HCC Computing Cluster Topics (NSF)

- Collaborative systems that enable knowledge-intensive and dynamic interactions for innovation and knowledge generation across organizational boundaries, national borders, and professional fields.

- Methods to support and enhance social interaction
  - Affective computing
  - Experience capture

- Social dynamics modeling and socially aware systems.
Cultural Factors in HCC

- Bowers (2000):
  - computers are not culturally neutral
  - current computing assumptions & practices have negative impacts on education, and preservation of cultures and the environment
  - other computing assumptions and practices (e.g. culture-specific) are possible
Cultural Factors not just in HCC

Zürich: a city and its trams
Cultural Factors in HCC
Social Factors in HCC

- **social interaction**
- example: dominance & influence
  - high-status people receive more visual attention than low-status (Efran, 1968)
  - people who rarely look at others while talking are perceived as weak (Cook, 1975)
  - high-status people exhibit a high ratio of looking-while-speaking to looking-while-listening periods (Exline, 1975)
  - people can decode patterns of visual dominance shown by others (Dovidio, 1982)
Experience

- **Device**: Mobile phones
- **Main applications**:
  - Experience management
  - Experiential communication
- **Users**:
  - Humans
  - No language issues
What Is a Human-centered System?

- A system that involves any human activity
  - Multimedia indexing (humans use images and video)
  - Camera-based Human-computer Interaction
  - Understanding of any sensory perceivable actions (e.g., eye, any body-part movement, emotions)

*and whose design uses human models or gives special consideration to human abilities*

- Utilize human memory, subjectivity, etc.

*We will focus on this set of problems...*
The Big Picture

- Detection
- Tracking
- Facial Expression
- Head/Body Pose
- Eye Location/Tracking
- Visual Gaze
- Applications
Living in Italy!!!!
Bridging the Emotional Gap:

Affective Multimedia Retrieval
Affective Multimedia Retrieval

Goal

- Extraction of affective content from multimedia data

Why is this important?

- Allows personalization of multimedia delivery and access

Why movies retrieval and sports video highlighting?

- One of the main applications in the consumer sector
Cognitive Retrieval Scenario

- **Examples:**
  - Find me
    - *all video clips containing a panorama of San Francisco!*
    - *all images of me and my Porsche taken in Paris!*
    - *all segments of my classical audio collection containing trills!*

- **Search for facts!**
What about other possibilities?

- Search for a subset of **nicest** holiday pictures to show them to friends
- Selecting **the most appropriate** background music for the given situation
- Search for **the most impressive** video clips
- Search for the **most appealing** photographs of one and the same content
- Search for all film comedies **I like most**

Real multimedia retrieval applications?
Alternative Retrieval Scenario

- Search for
  - Moods
  - Matches to user’s profile
    (like/dislike, interest/no interest)
Why “Affective”? 

- **Affect**: 
  - “a feeling or emotion as distinguished from cognition, thought, or action”

- **Affective state**: 
  - Specific emotional setting

- **Assumption**: 
  - Affect determines what the user likes or dislikes, what he is interested or not interested in
Cognitive versus Affective Multimedia Retrieval

- **Cognitive retrieval**
  - Made possible by establishing the relation between facts and low-level features
  - Bridging the semantic gap

- **Affective retrieval**
  - Made possible by expressing moods and user’s personal preferences using low-level features
  - Bridging the emotional gap
Dimensional Approach to Affect

- Affect has three basic dimensions
  - Arousal (intensity)
  - Valence (sign)
  - Dominance (degree of control)

- 3D emotion space
  - “parabolic” shape
  - obtained through psychophysiological experiments

Courtesy of Diets and Lang, Affective Agents, 1999
Creating a 3D emotion space

- Arousal = \( f \) (skin conductance)
- Valence = \( f \) (heart rate acceleration)
- Dominance - through self-report

Measured by affective reactions of large group of test users to sets of calibrated audiovisual stimuli

- IAPS (International Affective Picture System)
- IADS (International Affective Digitized Sound system)
From 3D to 2D emotion space

- Value range of dominance is very small
- Valence and arousal are known to determine most of affective states
- Besides:
  - Valence and Arousal are known to be fully sufficient for assessing TV broadcast regarding viewer’s interest!

Only the 2D Valence-Arousal space considered!

Courtesy of Diets and Lang, *Æffective Agents*, 1999
Affective Media Content Characterization

Media \xrightarrow{\text{Feature extraction}} A = f1 (feature values) 
\quad V = f2 (feature values)

Mapping AV values onto the 2D affect space

Affective content characterization
Matching with User’s Preferences

Feature extraction

A = f1 (feature values)
V = f2 (feature values)

Learning user’s preferences

Training Media

User

Comparing two areas

Match?

New Media

Feature extraction

A = f1 (feature values)
V = f2 (feature values)

Affective properties of user’s preferences

Unpleasant

Calm

Arousal

Unpleasant

Calm

Aroused

Pleasant

Affective media content properties
Case Study:

Affective Video Retrieval

Hanjalic & Xu
Motivation

- **Personalization of video delivery services**
  - possible using video analysis at affective level
  - no affective information sent by service provider

- **Meeting the affective retrieval requests**
  - Seems to be the main retrieval application in the consumer sector

- **Video highlighting**
  - Big business in media industry
  - Can be performed remotely
The underlying idea

- **Video**
  - Temporal content flow
  - Continuous transitions from one affective state to another

- **Temporal measurement of**
  - Arousal
  - Valence

- **Combining the two curves into the Affect Curve**
The Arousal and Valence Curve

- “most exciting” segments
- segments evoking “positive” or “negative” moods

“Most exciting segments”

Segments with a strongly “negative” mood
Affect Curve in the 2D Affect Space

Selection of all video segments with high arousal: Suitable to serving as video highlights

Selection of all video segments with dominating affective content: Suitable to sub-genre determination
Affect Curve and Affective Retrieval Applications

- Affective curve serves as the **matching criterion** between
  - user’s preference
  - user affective request
  and a video!
- Video highlights
  - parts of a video where the affect curve passes through upper left or upper right area of the 2D emotion space
Video database

Algorithm extracts affective content from the initial selection

USER PREFERENCES

User selects a number of video clips he likes most
User selects the “mood” he wants to find in a clip

Filtering according to user preferences

Personal program selection

Broadcast channels

Video highlights

Video database

Video IN

Algorithms extract affective content from incoming video

Broadcast channels
Major Challenges

- Computing a reliable affect curve for given video
  - Satisfying basic criteria such as
    - Comparability
    - Compatibility with 2D VA emotion space
    - Smoothness
    - Continuity
      \[ \text{Due to INERTIA of human feelings and emotions} \]
  - Identifying suitable low-level features

Finding ways of reliably comparing the affect curves of different videos
Comparing two Affect Curves

- Global comparison required
  - Based on the dominant curve behavior

Region of dominant curve behavior
Suitable low-level features for measuring arousal and valence:

- Resulting from psychophysiological experiments
  - Speech:
    - pitch, loudness, spectral energy in high frequencies, speech rate and intensity, rhythm, duration of last sentence’s syllable, voice quality (resonance)
  - Motion activity
- Intuitive
  - Density of cuts
  - Pitch - strongest indication to speaker’s intonation and stress
Modeling of Arousal

- Arousal is modeled as weighted average of three components:
  - motion activity
  - density of cuts
  - sound energy in higher frequencies

- Each component is computed as curve along a video

- Weighted average results in AROUSAL CURVE
Motion Activity (1)

- Step 1:
  - Frame-to-frame motion estimation
- Step 2:
  - Averaging of all motion vector magnitudes
- Step 3:
  - Normalization using maximum possible motion vector magnitude
- Result:
  - “Raw” motion activity curve $m(k)$
Motion Activity (2)

**Problem:**
Sharp peaks do not comply with inertia of human perception of arousal!
Motion Activity (3)

- **Step 4:**
  - Convolution of the $m(k)$ curve with a long Kaiser window and re-scaling into the same range as $m(k)$.
Arousal components

Motion activity

Density of cuts

Sound energy

Convolution

Slight smoothing
All components are scaled to the same range and are without dimension

⇒ Arousal value is independent on sequence-specific characteristics

⇒ Common base for comparison of arousal curves of different sequences
Modeling of Valence

- Also possible as weighted average of several components, like, e.g.
  - pitch
  - rhythm
  - voice quality (resonance)
  - ...

- Each component is computed as curve along a video

- Weighted average results in **VALENCE CURVE**

- In the following:
  - Development of a valence model based on **transposed pitch**!
The Transposed Pitch (1)

- Dividing the audio stream of a program into equal segments
- Averaging all pitch values within a segment that are pre-selected in the first step
- Assigning the average pitch value to all frames within a segment
Pitch Modeling: Example
The Transposed Pitch (2)

- A pitch value is reduced by a “neutral feeling” frequency
- Usual range of pitch values is transposed to a range circumventing zero line (neutral feeling line)!
- Convolution with a long Kaiser window to smooth vertical edges and re-scaling into the same range as the original pitch
Experimentally Obtained Arousal Curve

Average arousal: 4.8

Sequence: Segment from a football match
Length: 15 minutes

Average arousal: 8.29
Experimentally Obtained Valence Curve

![Graph showing valence curve with a "Neutral feeling" line.](image-url)
Experimental Results: Affect Curve

![Affect Curve Graph]

Movie
Affect Curve and Movie: Is there a link?
User Profiling

Feature extraction

Learning user’s preferences

A = f1 (feature values)
V = f2 (feature values)

Video

Feature extraction

A = f1 (feature values)
V = f2 (feature values)

Degree of matching

Preferred affective properties

Preferred affective properties

Preferred affective properties

Preferred affective properties

Unpleasant

Calm

Unpleasant

Calm

Pleasant

Aroused

Pleasant

Aroused

Preferred affective properties

Preferred affective properties

Preferred affective properties

Preferred affective properties

Affective properties

Affective properties

Affective properties

Affective properties
Personalized Affective Movie Delivery

List of programs already known to the system and having similar prevailing mood

User moves the pointer and scans the affective space
Other alternative?

Looking at the viewer!!!!
Multimodal Human-Computer Interaction:

Emotion Recognition
Affect-Sensitive HCI: Applications

- Research on natural (human-like) HCI: interfaces
- Video surveillance: behavior detection and matching
- Automatic assessment of boredom, stress, inattention in crucial but tedious tasks like air traffic control, driving a car, train, nuclear power plant surveillance
- Automatic affect-based indexing of digital visual material: video highlighting and content filtering
Emotions: What the Users Want?

“I have a nice photo of my darling and I want the computer to tell me if he still loves me!!!!”

Issues:
- Stimuli – Can you infer “love” from looking at a single photo?
- Context – Can I use the fact that I know that he is my lover?
- Dynamics – Would a static image be discriminative enough?
- Evaluation – Am I sure that the emotion displayed in the photo is really genuine? (Spontaneous vs. Posed)
- Multimodalities – Would some text written on the back of the photo help me to take a decision?
Emotion Evaluation: Stimuli

Two-factor Theories

- James’s Serial Account [1890]
  “We feel sorry because we cry, angry because we strike, afraid because we tremble.”

- Cannon’s Parallel Account [1927]
  Physiological responses and cognitive evaluations are independent effects of a common cause: the emotional stimulus.
Emotion Evaluation: Context

from Carlson & Hatfield, 1992
Emotion Evaluation: Dynamics

Single static

Multiple static

Dynamic

[Ambadar & Cohn04]
Emotion Evaluation: Dynamics

Single static

Multiple static

Dynamic
Affective Data Collection (Picard 2001)

- **Spontaneous** vs. **posed** (is the emotion elicited by an external situation or the subject is asked to elicit the emotion?)
- **Lab setting** vs. **real-world** (is the data recording in a lab or the emotion is recorded in the subject’s usual environment?)
- **Expression** vs. **feeling** (is the emphasis on external expression or on the internal feeling?)
- **Open recording** vs. **hidden recording** (is the subject aware that he is being recorded?)
- **Emotion-purpose** vs. **other-purpose** (does the subject know that he is a part of an experiment about emotion recording?)
Emotion Display: Multimodality

MICROPHONE

VIDEO CAMERA

ANIMATED AGENT

SYNTHETIC SPEECH

Speech Recognition
Vocal Affect
Head Movement
Facial Expression
Gesture Recognition
Eye Contact
Body Posture
Face Tracking

- Landmark facial features (e.g., eye corners) are detected automatically
- The generic face model is warped to fit those
- The face model consists of 16 surface patches embedded in Bezier volumes
- 2D image motions are measured using template matching between frames at different resolutions and 3D motion can be estimated from them
- The recovered motions are represented in terms of magnitudes of facial features
- Each feature motion corresponds to a simple deformation of the face
Facial Features

12 facial motion measurements

- vertical movement of the lips
- horizontal movement of the mouth corners
- vertical movement of the mouth corners
- vertical movement of the eye brows
- lifting of the cheeks
- blinking of the eyes
Facial Features

- We use 12 facial features = 12 facial motion measurements
- The combination of these features define the classes of facial expression: Neutral, Happy, Anger, Disgust, Fear, Sad, Surprise
Approaches

Static classification
• Uses feature vectors related to a single frame to perform classification
  • Bayesian Networks: Naïve Bayes & TAN (Tree-Augmented Network) (Cohen & Sebe, CVIU 2003)
  • Stochastic Structure Search (Semi-supervised learning) (Cohen & Sebe, PAMI 2004)

Dynamic classification
• Takes into account the temporal pattern in displaying the facial expression
  • Multi-level HMM classifier (Cohen & Sebe, CVIU 2003)
Beyond Naïve Bayes: TAN (Tree-augmented Networks) Classifiers

- There is an exact algorithm to find one-parent dependencies (trees) using Chow-Liu algorithm (1968).
- Unlike Naïve Bayes, the features are not independent given the class.
Gaussian TAN Classifier

Added complexity is small; one extra parameter to estimate per features (Covariance between pairs of features):

\[
p(f_i | \text{Class} = j, \text{parent}_i = x) \sim N(\mu_{ij} + a_{ij} x, \sigma_{ij}^2 (1 - \rho_{ij}^2))
\]

\[
a_{ij} = \frac{\text{Cov}(f_i, \text{parent}_i | j)}{\sigma_{\text{parent}_i}^2}, \quad \rho_{ij} = \frac{\text{Cov}(f_i, \text{parent}_i | j)}{\sigma_{\text{parent}_i} \sigma_{f_i}}
\]

[COHEN & SEBE – CVIU03]
Learning the TAN

= motion unit
Semi-supervised learning

- In our problem, labeled data is not easy to obtain compared to unlabeled data.
- It would be very beneficial if we could learn a good classifier with a small number of labeled data and an easily available large number of unlabeled data.
Theoretical Issues

- Unlabeled data always help reduce the estimator’s variance.
- When modeling assumptions are correct unlabeled data improve classification performance.
- Unlabeled data can degrade classification when modeling assumptions are incorrect; although the variance decreases, the estimation bias increases – THIS IS KEY!!!
Searching for structures

- The analysis suggests that we should pay attention to modeling assumptions when dealing with unlabeled data. Simple models such as Naïve Bayes or TAN are not good enough for our problem.

- In the context of Bayesian network classifiers, we must look for the correct graph structures

Idea: Search for Bayesian network structures using classification error as the search guiding score
Multi-level HMM: Lower-level

Face Tracking
Feature Extraction (MU)

Left-to-Right with return
or ergodic models

Decoded state
sequence vector
To top
level HMM

Happy Model
Sad Model
Surprise Model
Anger Model
Disgust Model
Fear Model
Multi-level HMM: Top-level

Output: State sequence of high-level HMM

Observations: State sequence of lower level HMMs

- Neutral
- Happy
- Sad
- Surprised
- Anger
- Fear
- Disgust

Observations:
- Sad
- Happy
- Neutral
- Surprised
- Anger
- Fear
- Disgust

State sequence of high-level HMM

Observations: State sequence of lower level HMMs
Multi-level HMM: Motivation

Using a two level HMM structure has two advantages over the one layered HMM:

- **Automatic segmentation of the video sequence**
  - The top-level HMM will learn the segmentation rule automatically (no heuristic rules)

- The use of the top-level will increase the discrimination between the output of the “expression-specific” HMMs
  - The top-level learns the “behavior” of each model for any expression sequence and takes it into account.
Chen-Huang Database: 5 persons, 6 sequences per emotion (separated by neutral)

For each person, one sequence of each emotion is left out, and the rest are used as training sequences

<table>
<thead>
<tr>
<th>Person</th>
<th>NB-Gauss</th>
<th>NB-Cauchy</th>
<th>TAN-Gauss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.97</td>
<td>81.69</td>
<td>85.94</td>
</tr>
<tr>
<td>2</td>
<td>87.09</td>
<td>84.54</td>
<td>89.39</td>
</tr>
<tr>
<td>3</td>
<td>69.06</td>
<td>71.74</td>
<td>71.78</td>
</tr>
<tr>
<td>4</td>
<td>82.5</td>
<td>83.05</td>
<td>86.58</td>
</tr>
<tr>
<td>5</td>
<td>77.18</td>
<td>79.25</td>
<td>82.84</td>
</tr>
<tr>
<td>Average</td>
<td>79.36</td>
<td>80.05</td>
<td>83.31</td>
</tr>
</tbody>
</table>

Emotion recognition rates accuracy (%)
## Results with Chen-Huang Database

<table>
<thead>
<tr>
<th>Subject</th>
<th>Standard HMM</th>
<th>Multilevel HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82.86%</td>
<td>80%</td>
</tr>
<tr>
<td>2</td>
<td>91.43%</td>
<td>85.71%</td>
</tr>
<tr>
<td>3</td>
<td>80.56%</td>
<td>80.56%</td>
</tr>
<tr>
<td>4</td>
<td>83.33%</td>
<td>88.89%</td>
</tr>
<tr>
<td>5</td>
<td>54.29%</td>
<td>77.14%</td>
</tr>
<tr>
<td>Average</td>
<td>78.49%</td>
<td>82.46%</td>
</tr>
</tbody>
</table>
Static vs. Dynamic Classification

- Dynamic classifiers (multilevel HMMs) are more suited for systems that are person-dependent
  - More sensitive to changes in appearance of expressions among different individuals
  - More sensitive to changes in temporal patterns
- Static classifiers are easier to train and implement
- Dynamic classifiers are more complex and need more training sample and many more parameters to learn
- When used for video sequences, the static classifiers can give unreliable results for frames outside the top of the expression
Emotion recognition

Play video
Static Images: Any Problems?

- There is no neutral expression available!!!
- However, we all show facial expressions in a consistent way (e.g., lift the eye brows when surprised, lower the corner of the mouth when sad)
- Create a statistical model from training data that represents an “average” neutral expression. Fit this model onto a particular face by adapting it to the characteristic features
Prototypical Neutral Face

White female

Black female
Mona Lisa’s Neutral Mask

Press the play button or press ↓ to fit the model
Fitting the Model

Press the play button or press ↑ to fit the model
Computing Displacements

Press the play button or press ↓ to fit the model

Press the play button or press ↓ to fit the model
Emotion Classification

Mood Recognition

Neutral 0.02%
Happy 83.13%
Surprise 0.0%
Angry 2.01%
Disgust 8.84%
Fear 5.87%
Sad 0.11%

Mood = Happy

Tracking
Mood Recognition

Neutral 0.0 %
Happy 0.05 %
Surprise 89.2 %
Angry 0.21 %
Disgust 0.0 %
Fear 10.25 %
Sad 0.28 %

Mood = Surprise
Bimodal Database

- 100 subjects
- 11 categories (6 basic plus interest, boredom, puzzlement, frustration and neutral)
- audio and video
- portrayal data
- more related to HCI
- non-content free environment: the subjects were asked to speak appropriate sentences for each categories (e.g., “computer, this is not what I asked you for!!!” for frustration)
Bimodal Database

Play video

Play video
Classification

Bayesian Network topology for recognizing emotions from audio and facial expressions (similar topologies can be used for other events)
Bimodal Database

(a) Happy expression without speaking; (b) Happy expression with speaking; (c) Surprise expression without speaking; (d) Surprise expression with speaking
Results

![Bar graph showing results for Face, Prosody, and Fusion categories with Male, Female, and No speaking categories.]
## Results

**Table 3. Male-female Average Confusion Matrix for the Bimodal Classifier**

<table>
<thead>
<tr>
<th>Desired</th>
<th>neut</th>
<th>hap</th>
<th>surp</th>
<th>ang</th>
<th>disg</th>
<th>fear</th>
<th>sad</th>
<th>frust</th>
<th>puzz</th>
<th>inter</th>
<th>bore</th>
</tr>
</thead>
<tbody>
<tr>
<td>neut</td>
<td>98.66</td>
<td>0.27</td>
<td>0.27</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.54</td>
<td>0.00</td>
<td>0.00</td>
<td>0.27</td>
<td>0.00</td>
</tr>
<tr>
<td>hap</td>
<td>4.07</td>
<td>89.15</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>1.36</td>
<td>0.00</td>
<td>0.00</td>
<td>1.02</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td>surp</td>
<td>4.70</td>
<td>0.31</td>
<td>88.40</td>
<td>1.57</td>
<td>0.00</td>
<td>1.57</td>
<td>1.88</td>
<td>0.00</td>
<td>0.94</td>
<td>0.00</td>
<td>0.63</td>
</tr>
<tr>
<td>ang</td>
<td>1.16</td>
<td>0.23</td>
<td>0.47</td>
<td>94.65</td>
<td>1.40</td>
<td>0.23</td>
<td>1.16</td>
<td>0.00</td>
<td>0.47</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>disg</td>
<td>2.31</td>
<td>0.77</td>
<td>0.77</td>
<td>3.08</td>
<td>88.72</td>
<td>0.77</td>
<td>0.77</td>
<td>1.03</td>
<td>0.51</td>
<td>0.51</td>
<td>0.77</td>
</tr>
<tr>
<td>fear</td>
<td>1.81</td>
<td>0.30</td>
<td>0.30</td>
<td>3.93</td>
<td>1.51</td>
<td>89.12</td>
<td>1.21</td>
<td>0.00</td>
<td>1.51</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>sad</td>
<td>4.60</td>
<td>0.31</td>
<td>0.31</td>
<td>3.07</td>
<td>1.53</td>
<td>0.00</td>
<td>88.04</td>
<td>0.61</td>
<td>0.31</td>
<td>0.31</td>
<td>0.92</td>
</tr>
<tr>
<td>frust</td>
<td>3.31</td>
<td>0.37</td>
<td>2.21</td>
<td>3.31</td>
<td>0.37</td>
<td>2.57</td>
<td>2.21</td>
<td>83.46</td>
<td>1.10</td>
<td>0.37</td>
<td>0.74</td>
</tr>
<tr>
<td>puzz</td>
<td>1.88</td>
<td>0.31</td>
<td>0.63</td>
<td>2.82</td>
<td>1.25</td>
<td>0.94</td>
<td>0.31</td>
<td>0.63</td>
<td>88.71</td>
<td>1.25</td>
<td>1.25</td>
</tr>
<tr>
<td>inter</td>
<td>2.14</td>
<td>0.36</td>
<td>1.42</td>
<td>3.20</td>
<td>1.78</td>
<td>1.07</td>
<td>1.42</td>
<td>0.36</td>
<td>1.07</td>
<td>85.05</td>
<td>2.14</td>
</tr>
<tr>
<td>bore</td>
<td>2.32</td>
<td>0.99</td>
<td>0.66</td>
<td>1.32</td>
<td>0.33</td>
<td>0.66</td>
<td>2.32</td>
<td>0.66</td>
<td>0.99</td>
<td>0.33</td>
<td>89.40</td>
</tr>
</tbody>
</table>
## Confusion Patterns

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Audio</strong></td>
<td>Hap ↔ Ang</td>
<td>Hap ↔ Ang</td>
</tr>
<tr>
<td></td>
<td>Sad ↔ Dis</td>
<td>Sad ↔ Dis</td>
</tr>
<tr>
<td></td>
<td>Ang ↔ Sur</td>
<td>Sur ↔ Fea</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ang → Sur</td>
</tr>
<tr>
<td><strong>Video</strong></td>
<td>Ang ↔ Dis</td>
<td>Ang ↔ Dis</td>
</tr>
<tr>
<td></td>
<td>Sad → Dis</td>
<td>Sad → Fea</td>
</tr>
<tr>
<td></td>
<td>Fea → Sur</td>
<td>Sur → Fea</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hap → Dis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sad → Sur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sad → Hap</td>
</tr>
<tr>
<td><strong>Audio/Video</strong></td>
<td>Ang ↔ Dis</td>
<td>Fea → Ang</td>
</tr>
<tr>
<td></td>
<td>Sad → Ang</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sad → Dis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fea → Sur</td>
<td></td>
</tr>
</tbody>
</table>
Questions to ask

- There exist “Universal Facial Expressions,” are there “universal vocal emotions?”
- How long a segment of data do we need before classifying the emotion?
- What are other modalities to explore?
- How can the multimodal approach be used in conjunction with single-modal methods?
- What are other ways of combining the modalities?
Automatic Assessment of Video Footage
Automatic Assessment of Video Footage

Neutral 0%
Happy 0%
Surprise 0%
Angry 0%
Disgust 0%
Fear 0%
Sad 0%

Status: Face Found!

Instructions: Press "Emotion" to analyse the face.
Footage Assessment
Product Assessment
Europe (Chocolate)
Italy (Ice cream)
Charisma in Politics: One killer app?
Visual Gaze Estimation

- Wearable systems are the most accurate
- Infrared systems are more accurate than appearance-based
- Non-wearable systems require calibration

Main issues
- Intrusiveness, speed, robustness, and accuracy
- Several applications (e.g., ambient intelligence) favor methods that do not require dedicated hardware
Visual Gaze Estimation
Pose vs Eyes

Eye Location

Head Pose
Eye Tracking

- Study the way people look at images
- Determine if there are fixation patterns
- Why is this important?
  - Understand visual process
  - Determine the interest of the user in a particular footage
  - Use regions of interest from eye tracking to interact with the computer
Cornea-Iris Detection
The iris and cornea are prominent circular features.

They are characterized by a quasi-constant intensity along the junctions sclera-iris and iris-cornea.

As such, these features can be represented using isophotes which are lines of equal intensity.
Isophote Properties

- **Isophotes:**
  - Don’t intersect each other
  - Are invariant to changes in contrast and brightness
  - Can fully describe a picture

- **Isophote Definition:**

\[
\hat{w} = \frac{\{L_x, L_y\}}{\sqrt{L_x^2 + L_y^2}} \;, \; \hat{v} = \perp \hat{w};
\]

\[
L(v, w(v)) = \text{constant}
\]
Isophote Curvature

- **Curvature** = change \( w'' \) of the vector \( w' \)
- By differentiating the definition of isophote:
  \[
  L_v + L_w w' = 0; \quad w' = -\frac{L_v}{L_w}.
  \]
- **Differentiating again:**
  \[
  L_{vv} + 2L_{vw} w' + L_{ww} w'^2 + L_w w'' = 0.
  \]
- \[
  \kappa = -\frac{L_{vv}}{L_w} = -\frac{L_y^2 L_{xx} - 2L_x L_{xy} L_y + L_x^2 L_{yy}}{(L_x^2 + L_y^2)^{3/2}}.
  \]
Isophote centers

- Now we have the curvature
  - Curvature is reciprocal of radius.
- Radius meaningless without orientation and direction
- Orientation from gradient

\[
\frac{\{L_x, L_y\}}{\sqrt{L_x^2 + L_y^2}} = \frac{\{L_x, L_y\}}{L_w}
\]
Isophote centers

- **Problem: Direction?**
- **Observation 1:** The sign of the isophote curvature depends on the intensity on outer side of the curve
  - Positive for light, negative for dark
- **Observation 2:** The direction of the gradient goes from dark to light
- **Solution:** Multiplication!
Isophote centers

\[
\text{isocenters} = \frac{\{L_x, L_y\}}{L_w} \left( -\frac{L_w}{L_{uu}} \right) = -\frac{\{L_x, L_y\}}{L_{xx}} \\
= -\frac{\{L_x, L_y\}(L_x^2 + L_y^2)}{(L_y^2L_{xx} - 2L_xL_{xy}L_y + L_x^2L_{yy})}
\]
Center voting through Curvedness

selection of isophotes is needed in real-life
Center voting through Curvedness

- Selection of isophotes using

\[ \text{curvedness} = \sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2}. \]
Eye center location

Algorithm 1 Pseudo-code for estimating isocenters.

- Compute first order, second order and mixed image derivatives $L_x, L_{xx}, L_{xy}, L_y, L_{yy}$
- Compute curvedness $= \sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2}$
- Compute $D = -\frac{L_{xx}^2 + L_{yy}^2}{L_y L_{xx} - 2L_x L_{xy} L_y + L_x^2 L_{yy}}$
  
  $Dx = L_x \times D$
  
  $Dy = L_y \times D$

Initialize an empty centermap
for $i = 1$ to image width do
  for $j = 1$ to image height do
    $x = i + \text{round}(Dx(i, j))$
    $y = j + \text{round}(Dy(i, j))$
    if $D > 0$ (for dark centers only) then
      $\text{centermap}(x, y) + = \text{curvedness}(i, j)$
    end if
  end for
end for
- Convolve centermap with a Gaussian kernel to cluster votes
Eye center location: Scale space
Pose – Normalized Visual Gaze
Pose vs Eyes

Head Pose

Eye Location
Motivation

- **Problems:**
  - eye detector only works on semi frontal faces
  - Head tracker can lose the correct track (stuck on local minima)
  - Head tracker cannot reinitialize when lost

- **Solution:**
  - Synergetic Eye Location and CHM Tracking!

- Deeper integration
  - eye location detected given the pose
  - pose adjusted given eye location
Integration of the Eye Locator

- Detect a frontal face (Viola & Jones)
- Use eye tracker to initialize the current transformation matrix
- Pass transformation matrix to the 3D model
- Use cylinder to extract/rebuild eye regions
- Start from 2

- Build a distribution map on the cylinder
- Calculate displacement from the mean
- Compensate difference between pose and eye gaze
- Find intersection with camera plane and match it

CVPR 2009, TIP 2012
Start

Initialization

Grab Frame

Face detected?

N

Detect eyes

Initialize cylinder by eyes

Y

Tracking

Grab next frame

Eyes detected?

N

Track head pose

Eyes locations by pose

Y

Track head pose

Eyes locations by pose

Update

Calculate error between pose vector of eye locations

Error > Threshold?

N

Update transformation matrix

Y
Rigid Pose Tracking – Eyes
MIC Operator - Interest
Saliency by Isocentric Curvedness and Color

Figure 4.1: An example of the conspicuity maps and their combination: (a) Curvedness, (b) Color boosting, (c) Isocenters clustering, (d) Combined saliency map, (e) Area with highest energy in the saliency map (red: detection, blue: ground truth).
Figure 4.3: An example of the obtained results. From top to bottom: Saliency map, graph-cut segmentation, segmentation weighted by saliency, ESS result (red: detection, blue: ground truth).
Saliency corrected gaze estimation
Saliency corrected gaze estimation
Saliency corrected gaze

Play video
within MIAUCE project
Applications: Web browsing

Play video
Applications: Driver Monitoring
Applications: Driver Fatigue
Summary and Future Research Directions
HCC Challenges

- Can we strike a good balance between user customization and automatic techniques?

- Speed, robustness, and reliability continue to be crucial issues.

- Can we change the way users interact with computers?
Outlook

● Multimodal communication is hard, even for humans
  – Good communication skills are good for your career! (if we ask this of humans, we want it from computers too!)

● Multimodal information is *often* more valuable than unimodal information
  – So is Multimodal HCC!

● What will drive it?
  – Lower costs, faster hardware
  – Mobile & ubiquitous computing
  – Globalization?
  – Smarter users
Questions

- Researchers focus on one modality (vision researcher, speech researcher). But MMHCC is multidisciplinary (psychology, sociology, etc.)

- How do we take advantage of the ubiquitous revolution? (vision is hard, we’re likely to do better in some problems by fusing input from other types of sensors)
Future Research Directions

- Develop theories based on human models
- Derive generic statistical models for human behavior
- Develop new, natural interaction techniques
- Do not ignore social impact and increased integration of technology on daily life
- Do not interfere with the user’s privacy
- Cultural differences can have a strong impact
Vision: Challenges & Opportunities

- **Generic features:** Can become unavailable or uninteresting
  - Opportunistic features
- **Frame scope:** Frames can have zero or misleading information value
  - Flexible fusion window
- **Calibration:** User-dependent function
  - Observation-based self calibration methods
- **Point-to-point accuracy metric:** May not be relevant to end application
  - Application-driven metric
- **Environment definition:** Manual, need to repeat upon movement
  - Observations of user interaction
- **Fixed vision function:** To avoid overloading the processor in real-time app.
  - Alg. switching, Task assignment (active vision)
- **Uniform value for data:** Misleading data can bias results
  - Confidence level based on content and history

Enabled by distributed sensing and processing, interfacing with higher layers
Vision: Challenges & Opportunities

- **Multi-camera vision:**
  - Redundancy, data fusion, role selection
  - Privacy management

- **In vision processing:**
  - Employ complementary methods
  - Associate confidence with data
  - Use context (incl. multimodal sensing) to guide vision

- **Address network deployment challenges:**
  - Online calibration
  - Calibration-free methods
  - Automated setup and environment discovery

- **Interfacing vision with reasoning modules**
Modeling Behaviors

- **User model**: (routines / preferences)
- **Environment discovery**: (objects, temporal/affine usage relations)
- **Camera network model**: (geometry, topology, roles)
- **Abnormal event detection**: (action, accident, zones, crowds)

**Behavior models**

**Behavior**

**Context**

**Events**

**Adaptation**

**High-level reasoning**

**Observation validation**

**Knowledge accumulation**

**Vision**

- Pose / activity
- Face / gaze
- User interface
- 3D model

**Vision hardware & network**
New Applications & Interfaces

Smart environments
social networks
ambient intelligence

behavior models
communication modes

high-level reasoning
visualization

behavior
context
adaptation
video / select view
voxel reconstruction

events
adaptation
avatar mapping
meta data

context
adaptation

behavior
context
adaptation

knowledge accumulation
observation validation
user preferences and availability

knowledge accumulation
observation validation
user preferences and availability

pose / activity
face / gaze
user interface
3D model

vision
multi-camera hardware & network
New Frontiers

- Design methodology for multiple applications on the same system
  - Multi-purpose smart spaces
- Fixed + mobile camera networks
  - Calibration-free methods
- How to model semantics (use of observations, history, user input, internet)
  - Ontology, knowledge representation, generalization
- Human-in-the-loop methodologies
  - Level / type of user interaction and query
  - “Logical” user vs. discovery based on “experimenter” user
New Frontiers

- Systems with adaptive behavior
  - Exploit “system ↔ user” adaptation in design
- Event modeling:
  - Adaptive granularity in space, time (when to declare an event, levels of description)
- Behavior change:
  - As a result of working with an observation (vision) system
  - As a result of working with an adaptive system
  - How to influence and measure it?